Automatic generation of pictorial transcripts of video programs

Behzad Shahraray & *David C. Gibbon*

Machine Perception Research Department
AT&T Bell Laboratories
Holmdel, NJ 07733-3030

ABSTRACT

An automatic authoring system for the generation of pictorial transcripts of video programs which are accompanied by closed caption information is presented. A number of key frames, each of which represents the visual information in a segment of the video (i.e., a scene), are selected automatically by performing a content-based sampling of the video program. The textual information is recovered from the closed caption signal and is initially segmented based on its implied temporal relationship with the video segments. The text segmentation boundaries are then adjusted, based on lexical analysis and/or caption control information, to account for synchronization errors due to possible delays in the detection of scene boundaries or the transmission of the caption information. The closed caption text is further refined through linguistic processing for conversion to lower-case with correct capitalization. The key frames and the related text generate a compact multimedia presentation of the contents of the video program which lends itself to efficient storage and transmission. This compact representation can be viewed on a computer screen, or used to generate the input to a commercial text processing package to generate a printed version of the program.

1. INTRODUCTION

The advent of digital multimedia communications has resulted in a considerable amount of activity in different aspects of multimedia computing and networking. Some of these efforts have been aimed at reducing the high storage and bandwidth demands of video and audio. Different compression methods resulting from these efforts have enabled the acquisition, storage, editing, and replay of video programs on reasonably powerful multimedia workstations. Other efforts have been concerned with the creation of models, methods, and tools for composing multimedia presentations. In using these tools, it has become evident that composing even a trivial multimedia presentation is a complex task and requires considerable time and effort on the part of the author. It is expected that the ongoing efforts will result in improved authoring methods which will minimize the involvement of the author in low level details related to timing and synchronization. However, when composing a multimedia presentation from scratch, the user of the system will continue to be responsible for specifying the structure of the presentation and the relationships between different media. In this paper, we discuss a special case of multimedia composition which simplifies the process considerably. This special case is that of converting an already composed multimedia presentation from one form, or one presentation environment, to another. This reduces the burden of authoring, and may even result in complete automation of the process by taking advantage of the time and effort that has been already spent in the production of the original program. More specifically, we are concerned with the generation of a compact, printable version of a video program.

Video programs are one form of a multimedia presentation consisting of at least two media; full motion video, and audio. They require a presentation environment which is capable decoding and displaying motion video, and playing audio. Printed material consisting of text and images, on the other hand, are self-contained and do not require any special purpose device (perhaps with the exception of spectacles!) to be perceived. As a result, printed material such as newspapers, books, and magazines are a major medium for information transfer. The process of converting a video program to a printed document which conveys the same information involves the selection of a small set of static images from the large number of video frames. This is achieved by performing a content-based sampling of the video frames. The text component may be provided by speech-to-text conversion. The current state of the art in automatic speech-to-text conversion, however, is below the level required for reliable conversion between the two media. Manual speech-to-text conversion by a human operator is a tedious task. However, we can take advantage of the effort that has already been spent to incorporate the closed caption signal into the video program. Many television broadcasts, such as news